Stochastic Difference Equations As Empirical Models of Population Fluctuations

Robert W. Poole

Biometrics Unit

Warren Hall

Cornell University

Ithaca, New York 14850

Submitted to Ecology

Running Head: Empirical Population Models

Abstract

The two purposes of this paper are to discuss some of the practical problems encountered in predicting fluctuations in population abundance and to suggest a specific class of empirical models of changes in population density called stochastic difference equations. Two distinct goals in population ecology are recognized; the discovery and description of how and why populations fluctuate and the prediction and control of the fluctuations of natural populations. Based on this distinction in goals, population models are divided into two classes; "What if" and "What will" models. A "What if" model asks "What happens if the causal variables and parameters take certain values?" A "What will" model asks "What will happen to a natural population for which the causal variables cannot be perfectly controlled?" Several practical problems make it difficult or impossible to use "What if" models to predict the fluctuations of natural populations. These practical problems include, 1) parameter estimation, 2) variable measurement, selection, and control, 3) sampling, 4) the complexity of ecosystems, 5) the importance of chance events and unknown variables, and 6) philosophical differences between mathematics and field biology.

The practical limitations of field ecology justify the use of empirical "What will" models. A useful model of population fluctuation should, 1) employ only measurable variables and estimable parameters, 2) be simple and intuitively appealing, 3) have applicability to a wide variety of problems, and 4) be stochastic. Two particular types of stochastic difference equations, autoregressive and autoregressive-moving averages, are studied and their properties, estimation, and identification are discussed. The models are used to predict and simulate fluctuations in the abundances of three species of Drosophila. The autoregressive models were judged superior to the autoregressive-moving averages models with respect to the data treated in this paper. Extensions of the models to other situations such as the modeling of fluctuations in communities of species and population control are discussed.

The two primary purposes of the population ecologist are to; 1) describe the pattern of interactions between species of the community and the physical environment and show how each factor of the environment directly or indirectly influences the changes in abundance of a population, and to 2) predict the fluctuations of the population and develop control schemes to maximize abundance, minimize abundance, or minimize the intensity of the fluctuations in abundance. Ecologists often fail to distinguish between these two very different and distinct goals. The methods and models used in achieving the first purpose are not necessarily the same as those needed to achieve the second. It may be possible to formulate a model of the fluctuations of an intensively studied population with variables and parameters with direct biological meaning. This type of model fulfills the first purpose of the ecologist and may be employed to study the effects of changing the levels of the parameters and variables of the model on the abundance of the population. This type of model may be termed a "What if" model. A "What if" model asks "What happens if the parameters and variables take certain values?" In contrast a "What will" model asks "What is going to happen in a real situation?" The simplicity of the "What if" model is that we are not required to estimate the parameters of measure the variables of the model. In many fields such as industrial engineering the experimenter has a great deal of control over the variables and parameters of a process. The experimenter can predict what will happen because he can control the levels of the variables and parameters of the process. In this type of situation "What if" and "What will" models are synonymous.

Field ecologists, unfortunately, rarely have much control over the important variables or parameters causing fluctuations in population abundance and often need to predict not only what can happen but also what will happen, purpose number two. The different purposes of the predictions of "What if" and "What will" models are usually overlooked in modeling the fluctuations of a population.

A question of even more generality is "What is reality in a model?" Models are abstractions and have no intrinsic reality beyond their ability or inability to fulfill the purposes they were created for. "What if" prediction and "What will" prediction are two distinct goals in ecology. The models used to achieve one goal are not necessarily or even likely to be the same models used to achieve the second. A model does not have to be biologically "real" to be useful predictively. The applied ecologist is certainly interested in understanding how his population responds to the environment, but more importantly he must predict the abundance fluctuations of his population. The applied ecologist needs a simple model providing good "What will" predictions and utilizing a minimum amount of easily gathered data.

The theoretically optimal model of a population or community of species populations would satisfy both purposes, but unfortunately there are a large number of practical problems mitigating against the creation of such saintly models. These problems are discussed below. In fact the difficulties encountered in working with uncontrolled field populations make it very unlikely that biologically "real" models will ever be used for the practical "What will" prediction of changes in population abundance. Empirical "What will" models of population change may be not only useful but absolutely necessary.

The dual purpose of this paper is to demonstrate the need for empirical "What will" models to predict the fluctuations in population of a species and to suggest a class of empirical models termed stochastic difference equations. The prediction problems the ecologist faces are far more analogous to the problems of the economist than of the physicist or engineer despite the emphasis in ecology on models and methods originally developed in the physical sciences. The complexity and unpredictability of economics has lead to the widespread use of empirically based models in economics and I believe empirical models have a purpose to serve in ecology as well.

Practical Problems in Population Modeling

The practical problems of predicting the changes in population abundance discussed below suggest that empirical "What will" models should posses the following properties:

1. The model should be useful predictively, employing only measureable variables and estimable parameters.
2. The model should be intuitively appealing and simple enough to be understood and used by the field biologist who is likely to need the model for prediction and control.
3. A single family of models should be sufficient for a wide variety of problems and purposes. If a single family of models applies to most of the problems a field biologist is likely to encounter, then the biologist has to learn only a single set of manipulations, calculations, and restrictions.
4. The models should be stochastic.

The practical problems of prediction motivating these four characteristics are discussed separately below.

Parameter estimation: The first important question to ask in formulating a model to predict population fluctuations is whether the parameters of the model can be estimated. In particular if the model is based on biologically meaningful characteristics of the population, can the parameters of this model such as birth, death, emmigration, and immigration rates be identified from the type of data it is feasible to gather? Unfortunately in many, if not most cases the answer is no. For example, mark-recapture methods are often used to estimate the losses and gains of individuals to a population. Individuals born in the study area and those entering the area by immigration are confounded, i.e. the birth and immigration rates of the population cannot be separated and are not identifiable. The same dilemma is true of death and emmigration rates. Theoretically the sampling program could be broadened to

gather the type of data needed to identify the parameters of the model. In practice, however, it is usually impossible to collect this type of data as any ecologist working with populations strongly influenced by dispersal well knows. We may be able to create a model based upon the biological characteristics of births, deaths, immigration, and emmigration, but in practice it may be very difficult, if not impossible, to estimate these parameters in natural populations.

If the variables of the model cannot be rigidly controlled, the biologist may well question the concept of the parameter itself. If reality is our criterion, then birth, death, immigration, and emmigration are not parameters, but stochastic variables changing with time and space. These variables are determined by other variables (including their own past histories) which are themselves influenced by another set of variables and so on ad infinitum. Even non-biologically defined parameters such as proportionality constants cannot be expected to remain unchanged in a natural system. Our choice of parameters depends upon the abstraction level of the model.

Once the level of abstraction of the model is choosen and a sampling program developed to gather the data needed to identify the parameters of the model, we are still left with the problem of finding statistically efficient and possibly unbiased methods of estimating these parameters. Parameter estimates with large variances or appreciable biases may seriously diminish the "What if" and "What will" predictions of a model. Unfortunately the statistical problems inherent in efficiently estimating the parameters of a model of a system as complicated as a population are far from trivial.

Variable measurement: If the purpose of a model is the "What will" prediction of changes in population abundance, a variable should not be included in the model if it is not measurable or is as inherently unpredictable as the fluctuations of the population. Suppose the abundance of a pest insect population were predicated to be

a function of the number of eggs laid per square meter the previous year. Counting the number of eggs per square meter in many insect species, however, is neither easy nor practical. Luckily it is usually possible, if still difficult, to estimate the number of eggs. The variance of the estimate of the variable can be very large in field problems however. If the predictions of the model are sensitive to the level of the variable, the estimation error may result in large deviations between the observed and predicted abundances of the population. In addition the estimation errors of the variable, the stochastic nature and non-independence of sequential observations of the variable, and the dependence of this variable on other factors of the environment can cause significant biases in the estimates of the parameters of the model.

A more serious problem in variable measurement is predicting some biological characteristic of a population from changes in the physical environment. Suppose the birth and death rates of a population are postulated to be caused in part by fluctuations in temperature and humidity. Models expressing birth and death rates as function of temperature and humidity are exceedingly valuable in determining why populations fluctuate, "What if" predictions. The expression of birth and death rates as functions of temperature and humidity for "What will" prediction is not feasible because fluctuations in temperature and humidity are as inherently unpredictable as the birth and death rates. In other words it is pragmatically impossible to predict the birth and death rates of a population at some future time $\underline{t}$ from the temperature and humidity at time $\underline{t}$ because temperature and humidity at time $\underline{t}$ are not known and are as difficult to estimate or guess at as the fluctuations in density of the population.

Complexity, unpredictability, and chance: It is a truism of almost absurd proportions to claim that the fluctuations of plant and animal populations are caused or influenced by an exceedingly large number of interrelated biotic and abiotic factors. Some of the

most commonly observed factors are changes in food supply or some other resource, the intrinsic biological characteristics of the species, the effect of the density of the population its rate of growth, immigration and emmigration, interactions with competing or interfering species, predators, parasitoids, parasites, diseases, fluctuations in abiotic factors particularly severe changes in importance factors of the physical environment such as rainfall, temperature, and so forth, and the actions of man. A species population is affected not only by itself and the direct action of other variables of the ecosystems, but also by the interactions between variables and the indirect effects of agents acting on the variables of the ecosystem determining the fluctuations of the population. Many of the factors determining changes in population abundance are inherently stochastic and unpredictable. There is nothing predictable about a cow stepping on a grasshopper, a seed falling on a rock instead of open ground, a gust of wind carrying a fly out to sea, or an ant getting hit on the head by a beer can thrown from a passing car. For example the species of goldenrod occurring in a study area may have different optimal adaptiveness to different environmental conditions and in a closed, homogeneous, small area, free of herbivores and closed to immigration and emmigration, one species will outcompete or exclude the other species given sufficient time (years). However, the environment of the real world is not constant or homogeneous, herbivores are always present, and areas are seldom closed to emmigration or immigration. In particular a major part in determining the species composition and abundances of an area is played by the dispersal or failure of dispersal of seeds of the different species, a factor controlled by the vagaries and whims of weather and luck. In this complex world of constantly changing environment, adaptiveness, and chance, the general principle of "Competitive Exclusion", although no less true, has little or no meaning. In fact, it is not uncommon to find seven or either species of goldenrods within a 20 foot circle and to observe over the years that the abundances of the species and the species

composition itself fluctuates in seemingly illogical and unpredictable ways.

Populations in the real world are usually influenced by a far larger number of interrelated and interacting variables than can be included in a model even if the effects of purely chance events are ignored. Chance events should not be ignored, however, because stochastic events may determine the eventual pattern of population fluctuations, particularly if the population is small at times. In fact many of the characteristics observed in the fluctuations of a population can only be explained in terms of stochastic events (see later in this paper). The actions of non-included variables and purely chance events lend a probabilistic component to every process whether this component is included in the model or not. If we want to predict population change in a realistic way, we must accept the fact that population fluctuation is an inherently stochastic process which can never be perfectly predicted. The logical answer to the unpredictability problem is to couch predictions in terms of the probabilities of possible outcomes. The purpose of stochastic models is probabilistic prediction. Stochastic models, however, are very difficult to work with analytically and only the simplest formulations of a model are practical to work with.

A Philosophical Problem: Perhaps the single greatest stumbling block in the use of mathematical models in ecology is the philosophical gap between the mathematician and the field biologist. The development of effective and biologically meaningful models requires a reasonably high level of mathematical sophistication, a sophistication generally impossible for the field biologist to attain and still have time to be a competent field biologist. The mathematician, on the other hand, may not understand the problems of the field biologist because a true appreciation of the realities and complexities of natural populations and ecosystems can only be gained by many years

of experience in the field and a fundamental knowledge of natural history. The field ecologist who understands the workings of the population or community is in no position to develop an effective mathematical model, and the mathematician often has little real understanding of the problem he is trying to solve. Even a "team" approach has its limitations because of the fundamentally different philosophies of the field biologist and the mathematician. The mathematician views his work in logical terms and extends his concept of logic to populations and ecosystems. The good field biologist, on the other hand, knows the real world of plants and animals and finds it anything but logical.

These practical problems and others suggest that a family of empirical models may be useful for "What will" predictions and should posses the four characteristics listed at the beginning of this section. One such family of models is the stochastic difference equation. The purposes of the remainder of this paper are to; 1) discuss the general properties, usefulness, restrictions, and biological interpretation of stochastic difference equations, 2) present relatively easy methods of estimating the parameters of these models, and 3) apply the models to data to determine the ability of each of the models to simulate and predict the fluctuations of the populations.

Stochastic Difference Equations

The dependence of the density of a population at time t on the past abundance of the population at t-1, t-2, t-3, and so forth is an important characteristic of populations which can be exploited in modeling population fluctuation. Suppose some measure of the abundance of the population is recorded at equal intervals of time, say weeks, months, or years. It should be emphasized that we will be dealing with an abundance and not true population density which is most often unmeasurable

in any case. Sequential measurements of the abundance of the population will in general be correlated, so we might specify that the abundance at time t, X_t, is a function of the past abundances of the population. The abundance observations are corrected for the mean to simplify parameter estimation so that $E(X_t)=0$. If X_t is postulated to be a linear function of the past abundance of the population, then

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \eta_t \qquad 1)$$

where the ϕ_k are the parameters of themodel, x_{t-k} the abundance of the population k time intervals ago from t, p the longest lag included in the model, and η_t an error term at time t of one sort or another. If the error terms η_t are random variables a_t such that $E(a_t)=0$, $E(a_t a_{t-k})=0$ $(k\neq 0)$, and $E(a_t^2)=\sigma_a^2$, i.e. the error terms have a constant variance, are independent, and have zero expectation, then Eq. 1 is termed an autoregressive equation.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t \qquad 2)$$

The error term η_t might also be specified to be generated by a moving averages process

$$\eta_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad 3)$$

where q is the longest lag included in the moving averages equation. Substituting Eq. 3 for η_t in Eq. 1 produces the autoregressive-moving averages model (ARMA)

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad 4)$$

The abundance of the population corrected for the mean is influenced at time t not only by the random error term a_t in an ARMA model, but also by a function of the previous error terms. Biologically the ARMA model may have a bit more meaning than the simple autoregressive model. In most practical problems the abundance of only one age group of a population, such as corn borer larvae or adult mosquitos, is of interest. The number of adult mosquitos at time t is postulated by the autoregressive model to be a linear function of the past abundances of adults and an error term. However, the number of adults at time t probably also depends on the effects of previous error terms a_{t-1}, a_{t-2}, and so forth on the survival of earlier age groups such as larvae, eggs, and pupae which give rise to the adults at time t. The importance of this biological interpretation of the ARMA model obviously depends upon the length of the time interval used and the biology of the species. It is important, however, not to attribute too much biological meaning to any of these models. The purpose of the models is solely to mimic the population fluctuations with the simplest model possible. Suppose the process of population fluctuation can best be represented by a moving averages model of small order, say $q=2$ or 3. A moving averages model is an ARMA model with $p=0$. Then equivalently the same process can be stipulated to be an infinite weighted sum of all of the past abundances of the population (Box and Jenkins, 1970). Conversely if the process is really a finite autoregressive process, the process can be written as an infinite weighted sum of all of the past, random error terms. There is no unique empirical model corresponding to the real fluctuations of the population because of the duality between the autoregressive and moving averages models. The stochastic difference euqtion model of population change has no intrinsic reality beyond its ability to predict changes in abundance and to simulate the general characteristics of the fluctuations of the population. The purpose of including both autoregressive and moving average terms in the model is to insure a parsimonious model. In most field problems, however, the abundance of a population is known to be functionally related to the past abundances of the population, and so

the model should contain at least an autoregressive component. The inclusion of a moving averages term in the model depends on whether the AR or the ARMA model model accounts for the largest percentage of the variance in X_t for a given number of time lags and parameters. The autoregressive model is preferable for practical reasons because the estimation of the parameters of the AR model and the use of the model predictively are easier than they are for the ARMA model.

Stationarity

Stochastic difference equations specify a stochastic process. If the model is to be useful predictively, this stochastic process should posses the property of stationarity. Stationarity implies that the joint probability distribution of a series of observations on the abundance of a population is not affected by shifts in time. More intuitively the mean, variance, and serial covariances of a series of observations on a weakly stationary process are not affected by changing the time period from which the samples are taken. Stationarity of the mean does not appear to be a serious restriction in the use of stochastic difference equations to predict population fluctuations. Most populations appear to fluctuate, sometimes severely, about a mean that might be loosely construed as the carrying capacity of the environment. Long term changes in the environment or the interactions of the population with other species can conceivably cause the mean level of the population to chance in a non-deterministic manner. Stationarity can usually be achieved in these cases by differencing the series of observations provided the behavior of the fluctuations remains homogeneous and only the mean is changing (see Box and Jenkins, 1970). Deterministic trends in the mean can also be corrected for (see Anderson, 1971).

One serious problem indirectly related to the stationarity of the variance of a process is the dependence of the error term a_t on the level of the process X_t. Extremely large fluctuations are often observed in insect populations and other species

strongly influenced by the physical environment and possesing scramble intraspecific competition. The potential variability of X_t is much smaller at low abundance levels than at high values of X_t because of the limit imposed on the population by zero, extinction, and because the magnitude of changes in X_t depends upon how many individuals there are to give birth to new individuals or to die. Therefore the error terms of the linear model depend in part on the level of abundance at time t in contrast to the assumption that $E(a_t^2) = \sigma_a^2$ for all t. This problem exists in the data analyzed in this paper. All abundances in this paper were therefore subjected to a $\log(x+1)$ transformation to remove, at least partially, the dependence of the variance on the level of the process before the correction for the mean was made and the data analyzed. The log transformation in effect specifies a non-linear model of a stochastic process which is less variable at low than at high population densities.

The Autoregressive Model

Let γ_s represent the serial covariance between the observations x_t and x_{t+s}

$$\gamma_s = E\left[(x_t - \mu)(x_{t+s} - \mu)\right]$$

Several estimators of γ_s have been proposed, but Box and Jenkins (1970) suggest

$$c_s = \frac{1}{N-s} \sum_{t=1}^{N-s} (x_t - \bar{x})(x_{t+s} - \bar{x})$$

where N is the number of observations available. The serial correlation between the abundance of the population at time t and $t+s$ is $\rho_s = \gamma_s / \gamma_0$, estimated as $r_s = c_s / c_0$. The series of autocorrelations ρ_0, ρ_1, ρ_2, and so forth is termed the autocorrelation function or correlogram. Similarly the series of serial

covariances can be termed the autocovariance function. Additionally $\gamma_{\underline{s}} = \gamma_{-\underline{s}}$.

Multiplying through the autoregressive equation (Eq. 2) by $x_{\underline{t}-\underline{s}}$

$$x_{t-s}x_t = \phi_1 x_{t-s}x_{t-1} + \phi_2 x_{t-s}x_{t-2} + \cdots + \phi_p x_{t-s}x_{t-p} + x_{t-s}a_t$$

and taking expectations results in the following relationship between the serial covariances

$$\gamma_s = \phi_1 \gamma_{s-1} + \phi_2 \gamma_{s-2} + \cdots + \phi_p \gamma_{s-p} \qquad 5)$$

Notice that the expectation of the product $x_{\underline{t}-\underline{s}}a_{\underline{t}}$ is zero because the abundance of the population at time $\underline{t}-\underline{s}$ does not depend on the error term $a_{\underline{t}}$ which is yet to happen at time $\underline{t}-\underline{s}$. Dividing Eq. 5 by $\gamma_0 = \sigma^2_{\underline{x}}$, the variance of the observations, results in

$$\rho_s = \phi_1 \rho_{s-1} + \phi_2 \rho_{s-2} + \cdots + \phi_p \rho_{s-p} \qquad 6)$$

Substituting for $\underline{s}=1,2,\cdots,\underline{p}$ in the above equation results in a series of $\underline{p}$ equations termed the Yule-Walker equations ($\rho_0=1$, $\rho_{\underline{s}} = \rho_{-\underline{s}}$)

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2} \\
&\vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p\rho_1
\end{aligned}$$

If we let $\underset{\sim}{\rho}_{\underline{p}}' = \left[\rho_1 \;\; \rho_2 \cdots \;\; \rho_p\right]$, $\underset{\sim}{\phi}_{\underline{p}}' = \left[\phi_1 \;\; \phi_2 \cdots \phi_p\right]$, and

$$\underset{\sim}{P}_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}$$

The Yule-Walker equations can be written more succinctly in matrix notation as

$$\underset{\sim}{\rho}_p = \underset{\sim}{P}_p \underset{\sim}{\phi}_p$$

If instead of the true but unknown autocorrelations we substitute in the estimated serial correlations and represent the estimates of $\underset{\sim}{P}_p$ and $\underset{\sim}{\rho}_p$ as $\underset{\sim}{R}_p$ and $\underset{\sim}{r}_p$, then estimates of the parameters of the model can be found by solving Eq. 7 for $\underset{\sim}{\phi}_p$, i.e. $\underset{\sim}{\phi}_p = \underset{\sim}{R}_p^{-1} \underset{\sim}{r}_p$.

Durbin (1960) has proposed a recursive method of estimating the parameters of the autoregressive model based upon the rule for the inversion of a partitioned matrix. The parameters of successively higher order models are found recursively from the next lower order equation. This recursive method of calculation has two advantages. With patience the calculations can be performed on a desk calculator and the order of the autoregressive model can be determined by successively increasing the order of the model unitl any further lag terms $\phi_k \underline{x}_{\underline{t}-\underline{k}}$ are effectively zero. For an autoregression of order one, i.e. $\underline{x}_{\underline{t}} = \phi_1 \underline{x}_{\underline{t}-1} + \underline{a}_{\underline{t}}$, $\phi_1 = r_1$. For a second order autoregression $\underline{x}_{\underline{t}} = \phi_1 \underline{x}_{\underline{t}-1} + \phi_2 \underline{x}_{\underline{t}-2} + \underline{a}_{\underline{t}}$

$$\phi_{21} = \frac{r_1(1 - r_2)}{1 - r_1^2}, \qquad \phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2}$$

where ϕ_{21} is the estimate of ϕ_1 of the second order model and ϕ_{22} the estimate of ϕ_2. For the third order model

$$\phi_{33} = \frac{r_3 - \phi_{21}r_2 - \phi_{22}r_1}{1 - \phi_{21}r_1 - \phi_{22}r_2}$$

$$\phi_{31} = \phi_{21} - \phi_{33}\phi_{22}$$

$$\phi_{32} = \phi_{22} - \phi_{33}\phi_{21}$$

In general

$$\phi_{p+1,p+1} = \frac{r_{p+1} - \sum_{j=1}^{p} \phi_{pj} r_{p+1-j}}{1 - \sum_{j=1}^{p} \phi_{pj} r_j} \qquad j=1,2,\cdots,p \qquad 8)$$

$$\phi_{p+1,j} = \phi_{p,j} - \phi_{p+1,p+1}\phi_{p,p-j+1} \qquad 9)$$

Successively higher order autoregressions are estimated by; 1) estimating $\phi_{p+1,p+1}$, the last term of the autoregressive model of order $p+1$ from Eq. 8 and, 2) modifying each of the parameters of the pth order autoregression by Eq. 9. The order of the model is increased until all higher values of $\phi_{p+1,p+1}$ are essentially zero. If the number of obeservations is reasonbly large, the quantity $N^{\frac{1}{2}}\phi_{pp}$ is approximately a standardized normal variable (Anderson, 1971). The series of values of $\phi_{p,p}$ is termed the partial autocorrelation function by analogy with partial correlation coefficients. An estimate of the variance of the error terms of the model is

$$s_a^2 = s_X^2(1 - r_1\phi_1 - r_2\phi_2 - \cdots - r_p\phi_p)$$

This estimate of the variance of the error terms may be though as a residual variance, and our goal is to find themodel with the smallest estimated error variance s_a^2.

The estimation of the parameters of the autoregressive model with the Yule-Walker equations is not as efficient as estimation with the least squares procedures discussed in Box and Jenkins (1970) or in Anderson (1971) and used by Poole (1972), but if the series of observations is not approaching the limits of stationarity (see below) the estimation procedure is reasonably good. The covariance matrix of the estimates is

$$\text{Var}(\underset{\sim}{\hat{\phi}}) = N^{-1}(1 - \underset{\sim}{r}'\underset{\sim}{\phi})\underset{\sim}{R}^{-1}$$

where $\underset{\sim}{r}$ and $\underset{\sim}{R}$ are sample estimates of $\underset{\sim}{\rho}$ and $\underset{\sim}{P}$. The standard errors of the estimates are the square roots of the diagonal elements of the matrix $\text{Var}(\underset{\sim}{\hat{\phi}})$ and the correlation between the estimates of two parameters ϕ_i and ϕ_j is

$$\text{Cor}(\hat{\phi}_i, \hat{\phi}_j) = \frac{\text{Cov}(\hat{\phi}_i, \hat{\phi}_j)}{(\text{var}\ \hat{\phi}_i\ \text{var}\ \hat{\phi}_j)^{\frac{1}{2}}} \qquad i \neq j$$

If a high negative or positive correlation exists between the estimates of two parameters ϕ_i and ϕ_j, the confidence region for ϕ_i and ϕ_j will be attenuated in one direction implying that the estimates of the parameters are quite unstable.

Hunter (1966) studied the fluctuations in abundance of several species of Drosophila in a government protected pine plantation near Bogotá, Colombia. The three commonnest species were Drosophila pseudoobscura, D. mesophragmatica, and D. viracochi. Hunter sampled the species at bait traps monthly from September, 1961, to December, 1963 for a total of 28 montly abundance figures. In Hunter's tables the commonness of each species is expressed as its percent frequency of the total number of flies. Also listed are the total number of traps used, total number of flies caught, and the average number of flies per trap. Because the number of traps at each site changed from month to month, the percentages were changed to number of flies based on the average number of flies per trap rather than the total number of flies. These abundance figures were subjected to a $\log(x+1)$ transformation, as

already mentioned, and the sample mean subtracted before the model was fit to the data. The abundance fluctuations of the three species on a $\log(x+1)$ scale are shown in Fig. 1.

The autocovariances and autocorrelations of the fluctuations in abundance of the three Drosophila species for lags of 0 to 10 are given in Table 1. Only 28 observations were available and it appeared useless to try to estimate autocorrelations and covariances of an order greater than ten because of the paucity of observations, the rapid loss of degrees of freedom, and the rapidly increasing variances of the estimates. Even the lower order estimates have relatively large variances because so few observations are available. The partial autocorrelation function, estimates of the error variance for each order model up to ten, and the terms $N^{\frac{1}{2}}\phi_{pp}$ are listed in Table 2.

The partial autocorrelation function of Drosophila mesophragmatica is the easiest to interpret. Lags of one and four are highly significant, but all higher lags up to ten are apparently non-significant. Therefore a fourth order model was choosen for mesophragmatica.

$$x_t = .87x_{t-1} - .36x_{t-2} + .46x_{t-3} - .53x_{t-4} + a_t$$

The estimated error variance of the fourth order model is .0603 versus a s_x^2 of .1719. The partial autocorrelation functions of pseudoobscura and viracochi are not as easily interpreted. The first four terms of the partial autocorrelation function of pseudoobscura are all relatively large, particularly the first and fourth. The fifth term is not significantly different from zero with 95 percent probability, but is still fairly large. The behavior of the function for lags of eight and higher is distressing because the estimation procedure is obviously going awry. The higher order matrices $\underset{\sim}{R}$ are probably ill-conditioned for one reason or another, perhaps because of the possible non-stationarity of the short

series of observations available. A fourth order model, however, appears to be appropriate for pseudoobscura.

$$x_t = .83x_{t-1} + .17x_{t-2} - .78x_{t-3} + .68x_{t-4} + a_t$$

The estimated error variance is $s_a^2 = .0380$ versus $s_x^2 = .1721$.

The appropriate order autoregressive model for Drosophila viracochi is difficult to determine. Although the sixth order lag is not significantly different from zero, the lag is fairly large. A fourth order equation is the logical model because a fourth order model appears to be appropriate for the other two species. Although the patterns of the fluctuations in abundance of the three species are different, the species are congeneric and it may be reasonable to assume that the dependence lags relevant to one species would be relevant to all. However, the sixth order equation is possibly a better model of the fluctuations in abundance of viracochi. Choosing a fourth order model as appropriate

$$x_t = .45x_{t-1} + .02x_{t-2} + .57x_{t-3} - .39x_{t-4} + a_t$$

with an estimated s_a^2 of .0715 versus $s_y^2 = .1362$. The fluctuations in abundance of of viracochi are not well predicted by the autoregressive model relative to the other two species by the error variance criterion. Inclusion of the sixth order lag in the model reduces the error variance to only .0612.

The covariance matrices of the parameter estimates for the fourth order models of the three species are listed in Table 3. The standard errors of the estimates are fairly large, but reasonable because the estimates are based on only 28 observations.

The parameters of the fourth order models were also estimated by the non-linear least squares procedure discussed in Box and Jenkins (1970) to determine how

well the Yule-Walker parameter estimates matched the more efficient least squares estimates. The Yule-Walker estimates were used as initial estimates and iteration terminated with convergence to four decimel places. The vectors of parameter estimates are (.81, -.33, .49, -.56; s_a^2 = .0752) for Drosophila mesophragmatica, (.68, .29, -.54, .41; s_a^2=.0745) for D. pseudoobscura, and (.38, .12, .62, -.42; s_a^2=.0856) for viracochi. The two sets of estimates compare well for mesophragmatica, fairly well for viracochi, and not too well for pseudoobscura. The Yule-Walker estimates of the fourth order model parameters of D. pseudoobscura are poor relative to the efficient least squares estimates because the fluctuations of pseudoobscura are probably non-stationary in the mean. In fact, the correlogram of pseudoobscura fails to damp to zero through the first ten lags indicating the presence of non-stationarity in the mean of this species. The correct procedure for pseudoobscura is to difference the observations, $z_t = x_t - x_{t-1}$, and work with the series of differences z_t. The estimates of the parameters of a fourth order model applied to the differenced data are well behaved, but only the non-differenced model is discussed for purposes of comparison.

Simulations of the fluctuations of the three species were conducted with the fourth order models to determine if the pattern of fluctuations of the three species are reasonably well mimicked by the autoregressive models. The fluctuation of a population is an inherently stochastic process and any agreement between the simulations and the data is qualitative only beyond the first few observations in each series. The simulations were performed by assuming that the error terms of the model are independently and normally distributed with zero mean and variance σ_a^2. Each simulation was started with the first four observations in the series. Simulations of mesophragmatica and pseudoobscura are shown in Fig. 2. On comparing Fig.2 and Fig. 1 Drosophila mesophragmatica appears to be the best mimicked of the three species. The other two species also appear to be reasonably well represented qualitatively, but not as well as mesophragmatica. The stochastic process generated by the autoregressive model of pseudoobscura was the most sensitive to the particular

set of random deviates used. The simulation in Fig. 2 resembles the original observations fairly cloosely, but other simulations are quite diverse in appearance.

Properties of the Autoregressive Model

The autocorrelation function of the stochastic process generated by an autoregressive model is given by Eq. 6. Define the backward operator $\underline{B}$ as $\underline{B}\rho_{\underline{s}} = \rho_{\underline{s}-1}$ and $\underline{B}^{k}\rho_{\underline{s}} = \rho_{\underline{s}-\underline{k}}$. Utilizing this operator notation Eq. 6 can be written as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\rho_s = 0$$

The general solution of this homogeneous difference equation depends upon the roots of the characteristic equation $(1 - \phi_1\underline{B} - \cdots - \phi_{\underline{p}}\underline{B}^{p}) = 0$. If a process is stationary, all of the roots of the characteristic equation must be greater than one in absolute value or modulus (Box and Jenkins, 1970). If a pair of roots is complex, the pair of complex roots contributes a damped sign wave to the autocorrelation function. A real root contributes a geometrically decaying term. Most autoregressive processes of an order higher than four consist of a mixture of both components. A process dominated by complex roots may exhibit pseudo-periodic behavior, i.e. the abundance of the population will fluctuate with a variable but defined period and amplitude.

The roots of the characteristic equations of the three fourth order autoregressive models applied to the Drosophila species are given in Table 4. All roots are greater than one in absolute value or modulus. Both mesophragmatica and viracochi have two pair of complex roots and should exhibit pseudo-periodic behavior (see below). In constrast the characteristic equation of pseudoobscura posses two real roots and one pair of complex roots.

The properties of the autoregressive models, and indirectly of the fluctuations of the populations, can be studied by calculating the spectral density function associated with the model. The power spectrum of a process for a frequency f and period $1/f$ is

$$p(f) = \frac{2\sigma_a^2}{\left|1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f} - \cdots - \phi_p e^{-i2p\pi f}\right|^2} \qquad 10)$$

where σ_a^2 is the error variance of the process, p the order of the model, and $i=(-1)^{\frac{1}{2}}$. The spectral density function $g(f)$ is $p(f)/\sigma_x^2$. The domain of f is 0 to $\frac{1}{2}$. The shortest period that can be detected is two time intervals or two months in this study. The squared modulus in Eq. 10 can be written as

$$(1 - \phi_1 \cos 2\pi f - \ldots - \phi_p \cos 2p\pi f)^2 + (-\phi_1 \sin 2\pi f - \cdots - \phi_p \sin 2p\pi f)^2$$

for computational purposes. The spectral density function is a measure of the density of fluctuations with frequency f and period $1/f$ and may be thought of intuitively as the frequency distribution of the time periods between the peaks or valleys of the fluctuations.

The spectral density functions of the fourth order autoregressive models applied to the three species of Drosophila are shown in Fig. 3. The autoregressive models associated with mesophragmatica and viracochi exhibit distinct pseudo-periodic behavior. The theoretical fluctuations of mesophragmatica are concentrated about a mean period of about 12 months. The theoretical fluctuations of viracochi are concentrated about a mean period of 20 months (24 months when the efficient least squares estimates are used). The majority of the variability of the periods of the fluctuations of mesophragmatica lies between 10 and 14 months. Drosophila mesophragmatica, therefore, appears to have a fairly well defined yearly cycle.

The peak in the spectral density function of viracochi may correspond to an average two year cycle in the fluctuations of this species. The cycle produced by the model is pseudo-periodic because although two years is the average period between peaks in abundance, the time period between peaks varies between 27 and 17 months.

The spectral density function of pseudoobscura is greatest about zero, i.e. the fluctuations produced by the model tend to be indefinite in length. There is, however, a slight increase in the density at $f=.05$ corresponding to a very weak and variable pseudo-periodic behavior averaging 20 months (24 months for the efficient least squares estimates), curiously the same average two year period as in viracochi.

Theoretically the adequacy of the models in reproducing the actual fluctuations of the population could be tested by comparing the theoretical spectrum of the model with the sample spectrum of the data (see Hannan, 1960). Unfortunately it appears futile to try to estimate the sample spectrum of the species fluctuations from only 28 observations.

The Autoregressive - Moving Averages Model

A more parsimonious representation of the fluctuations of a population could perhaps be achieved by including a moving averages term in the model. The autoregressive-moving averages model will be designated as ARMA(p,q) where p and q are the orders of the autoregressive and moving averages segments of the model respectively. Box and Jenkins (197) suggest the following method of finding preliminary estimates of the parameters of the ARMA model. The autoregressive parameters are first estimated by solving the p linear equations

$$
\begin{aligned}
C_{q+1} &= \phi_1 C_q + \phi_2 C_{q-1} + \cdots + \phi_p C_{q-p+1} \\
C_{q+2} &= \phi_1 C_{q+1} + \phi_2 C_q + \cdots + \phi_p C_{q-p+2} \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \\
C_{q+p} &= \phi_1 C_{q+p-1} + \phi_2 C_{q+p-2} + \cdots + \phi_p C_q
\end{aligned}
$$

for $\underset{\sim}{\phi}$ where C_q and so forth are the estimated autocovariances of order q and higher. Given the estimate of $\underset{\sim}{\phi}$, the autocovariance C'_j of the derived series

$$x'_t = x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \cdots - \phi_p x_{t-p}$$

are calculated as

$$C'_j = \sum_{i=0}^{p} \phi_i^2 C_j + \sum_{i=1}^{p} (\phi_0\phi_i + \phi_1\phi_{i+1} + \cdots + \phi_{p-i}\phi_p) d_j$$

where $j=0,1,\ldots,q$; $d_j = C_{j+1} + C_{j-1}$, and $\phi_0 = -1$. The error or residual variance and the moving averages parameters are estimated iteratively from the series C'_j

$$s_a^2 = \frac{C'_0}{1 + \theta_1^2 + \cdots + \theta_q^2}$$

$$\theta_j = -\left(\frac{C'_j}{s_a^2} - \theta_1\theta_{j+1} - \theta_2\theta_{j+2} - \cdots - \theta_{q-j}\theta_q \right)$$

For an ARMA(4,4) model

$$s_a^2 = \frac{C'_0}{1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \theta_4^2}, \qquad \theta_4 = -\left(\frac{C'_4}{s_a^2} \right)$$

$$\theta_3 = -\left(\frac{C'_3}{s_a^2} - \theta_1\theta_4 \right), \qquad \theta_2 = -\left(\frac{C'_2}{s_a^2} - \theta_1\theta_3 - \theta_2\theta_4 \right)$$

$$\theta_1 = -\left(\frac{C'_1}{s_a^2} - \theta_1\theta_2 - \theta_2\theta_3 - \theta_3\theta_4 \right)$$

The iterative procedure is started with initial values $s_a^2 = C'_0$, $\theta_1=\theta_2=\theta_3=\theta_4=0$, and proceeds from top to bottom using the newest values of the parameters available at

each step in the calculations. Convergence is linear and can be painfully slow. The theoretical power spectrum of an ARMA process is

$$p(f) = 2\sigma_a^2 \frac{\left|1 - \theta_1 e^{-2i\pi f} - \cdots - \theta_q e^{-i2q\pi f}\right|^2}{\left|1 - \phi_1 e^{-i2\pi f} - \cdots - \phi_p e^{-i2p\pi f}\right|^2} \qquad 0 < f \leq \tfrac{1}{2}$$

The parameters of ARMA(4,4) models were estimated for all three species. Ten to twenty iterations were needed to achieve convergence to two decimel places. In none of the three species did the ARMA(4,4) model represent an improvement over the AR(4) model. As an example the ARMA(4,4) model fitted to the pseudoobscura data resulted in the following equation:

$$x_t = 1.18x_{t-1} - .32x_{t-2} - .64x_{t-3} + .55x_{t-4} + a_t - .77a_{t-1} + .38a_{t-2}$$

$$+ .18a_{t-3} + .37a_{t-4}$$

The error variance associated with this model is .0619 compared with error variance .0380 for the AR(4) model with the Yule-Walker estimates. The autoregressive model contained only half as many parameters, had a smaller error variance, and was easier to estimate than the corresponding ARMA model if the comparison is based upon the preliminary parameter estimates.

The ARMA(4,4) model fitted to the fluctuations of Drosophila viracochi is the only model which might possibly be considered an improvement over the corresponding AR(4) model. The error variance associated with the ARMA(4,4) model is .0638 compared with .0715 for the AR(4) model. However, the reduction in the error variance achieved by including the four moving averages parameters in the model is small. In fact, the use of a sixth order AR model results in a reduction in the error variance to .0612 with only six parameters rather than the eight parameters of the ARMA(4,4) model.

The ARMA(4,4) model fitted to the mesophragmatica data was the most discouraging. The estimation procedure converged until the seventh iteration and then rapidly begin diverging. The comparison of the AR(4) and ARMA(4,4) models, however, is completely specious because when the efficient non-linear least squares estimation procedure is used for estimating the parameters of the ARMA(4,4) model, convergence does not occur in any of the three species. The only ARMA model estimates which will converge are the parameters in the ARMA(1,1) model, and the estimated error variance is in all cases considerably larger than the error variances of the AR(4) models.

The most probable cause of the failure to converge is the existence of nearly equal factors in the associated polynomial equations on the two sides of the equation. Suppose the ARMA model were

$$(1 - .5B)(1 - .8B)x_t = (1 - .5B)a_t$$

The redundancy of the factor $(1 - .5B)$ on both sides of the equation causes complete instability in the parameter estimates. In practice extreme instability results even by near cancellation of factors on the two sides of the equation. In other words the model is being overfit with too many parameters. The theoretical solution to this problem is to preliminarily identify the probable orders of the autoregressive and moving average segments of the model using methods given in Box and Jenkins (1970). Unfortunately these methods depend on matching the observed autocorrelation and partial autocorrelation functions of the series with the function produced by AR and ARMA models of low order, i.e. $p,q,=1,2$. The autocorrelation and partial autocorrelation functions of the fluctuations of the three Drosophila species are more complex than the primary types listed in Box and Jenkins (1970) indicating the necessity of higher order models and posing the problem of which order ARMA process is most appropriate.

The autoregressive-moving averages models did not provide better predictive or more parsimonious models than the corresponding autoregressive models. Conceivably, however, for other species the ARMA models would be more parsimonious of parameters and perhaps better predictively than the simpler autoregressive models.

Forecasting

The purpose of formulating an empirical model of the fluctuations of a population is to forecast the abundance of the population at some future time. The recursive form of the difference equation used in autoregressive and autoregressive-moving averages models makes prediction a relatively simple problem, i.e. the abundance one time interval in the future is forecast as

$$\hat{x}_{t+1} = \phi_1 x_t + \phi_2 x_{t-1} + \cdots + \phi_p x_{t-p+1} - \theta_1 a_t - \cdots - \theta_q a_{t-q+1}$$

Continuing this recursive procedure, the predicted abundance for a lead time of $l=1,2,\cdots,s$ is

$$\hat{x}_{t+l} = \phi_1 \hat{x}_{t+l-1} + \phi_2 \hat{x}_{t+l-2} + \cdots + \phi_p \hat{x}_{t+l-p} - \theta_1 \hat{a}_{t+l-1} - \cdots \theta_q \hat{a}_{t+l-q}$$

if $l \geq p,q$ or with the hats removed over known values of x_t and a_t is l is not greater than p or q or both. In other words future forecasts are calculated recursively from the past known or predicted values of the population abundances and error terms. The one remaining but most important problem is to estimate the variance or standard error of the predicted abundance at time $t+l$

A stationary stochastic process generated by an autoregressive or autoregressive-moving averages model can be represented by an infinite weighted sum of all of the past error terms (Box and Jenkins, 1970).

$$x_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \qquad 11)$$

$$= a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j}$$

Using operator notation

$$x_t = \psi(B) a_t \qquad 12)$$

However, the genral ARMA model can be written as $\phi(B)x_t = \theta(B)a_t$ and substituting Eq. 12 for x_t in this expression yields $\phi(B)\,\psi(B) = \theta(B)$. The ψ_i's, therefore, can be found by equating coefficients on both sides of this last equation. In particular

$$\begin{aligned}
\psi_1 &= \phi_1 - \theta_1 \\
\psi_2 &= \phi_1 \psi_1 + \phi_2 - \theta_2 \\
\psi_3 &= \phi_1 \psi_2 + \phi_2 \psi_1 + \phi_3 - \theta_3 \\
&\vdots \quad \vdots \qquad \vdots \qquad\qquad \vdots \\
\psi_j &= \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \cdots + \phi_p \psi_{j-p} - \theta_j
\end{aligned} \qquad 13)$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$. If the counter k is the greater of $p-1$ and q, then for $j > k$ the ψ's satisfy the differenc e equation

$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \cdots + \phi_p \psi_{j-p}$$

If the model is only an autoregressive process, the ψ's are calculated by simply dropping the moving average parameters θ_j from Eq. 13.

Returning to the prediction problem, Eq. 11 allows us to write x_{t+l} as

$$x_{t+l} = \sum_{j=0}^{\infty} \psi_j a_{t+l-j} \qquad 14)$$

Suppose, then, that the best prediction or forecast of the abundance of the population at time $t+l$ is generated by

$$x_t(l) = \psi^*_l a_t + \psi^*_{l+1} a_{t-1} + \psi^*_{l+2} a_{t-2} + \cdots \qquad 15)$$

where the ψ^*_{l+j} are unknown weights. The forecast error, $x_t(l) - x_t(l)$, from Eqs. 14 and 15 is

$$e_t(l) = \psi_0 a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + (\psi_l - \psi^*_l) a_t + (\psi_{l+1} - \psi^*_{l+1}) a_{t-1} + \cdots$$

Recalling that the a_t's are independently and identically distributed, the mean square error of the forecast is

$$E\left[e_t(l)^2\right] = (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 + \sum_{j=0} (\psi_{l+j} - \psi^*_{l+j})^2 \sigma_a^2$$

Clearly the mean square error is minimized if ψ^*_{l+j} is set equal to ψ_{l+j}. If we set ψ^*_{l+j} equal to ψ_{l+j}, the forecast error, i.e. the difference between the observed and predicted abundance at $t+l$ is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$$

Because $E\left[e_t(l)\right] = 0$, the forecast of the abundance of the population for a lead time l is unbiased with variance

$$\text{Var}\left[x_t(l)\right] = \left[1 + \sum_{j=1}^{l-1} \psi_j^2\right] \sigma_a^2$$

In practice this variance is estimated by replacing $\sigma^2_{\underline{a}}$ with its sample estimate $s^2_{\underline{a}}$. If we can assume that the $\underline{a}_t$'s are normally distributed, then an approximate $1-\alpha$ confidence interval is

$$x_t(l) \pm z_{\alpha/2}\left\{1 + \sum_{j=1}^{l-1} \psi_j^2\right\}^{\frac{1}{2}} s_a$$

where $z_{\alpha/2}$ is a standard normal deviate, .674 for a 50 percent and 1.96 for a 95 percent confidence interval.

The weights, predicted abundances, and variances of the estimates for the fourth order autoregressive models applied to each species were calculated for lead times of $\underline{l}=1,2,\cdots,10$. The predicted abundances of mesophragmatica and approximate 50 percent confidence intervals for lead times one to ten are shown in Fig. 4. In each case the recurrence relationship was begun with the first four observed abundance observations. Because the observations and predictions are on a log scale, the deviations from the predictions will be considerably larger at high abundances than at low abundances. As the lead time increases, the predictability of the population density decreases, as you would expect. Consequently the width of the confidence intervals increases with increasing $\underline{l}$. The sequential observations are not independent and if the predicted series falls above or below the confidence band, it is likely to stay there for some time.

Discussion and Conclusions

Stochastic difference equation models of population fluctuation, in spite of their obvious weaknesses, do seem to supply a simple and fairly reliable method of extracting the maximum amount of predictability from a minimum amount of data. These empirical equations are static models and their primary purpose is prediction. However, stochastic difference equations can be extended to a wide variety of other

problems, including the development of dynamic models and control schemes. If measurements on a non-stochastic, controllable variable are available, such as levels of pesticide applications, these non-stochastic variables can be included in the autoregressive model without significantly changing the maximum likelihood or least squares estimation of the parameters of the model (Anderson, 1971). Non-stationarity of the mean, either seasonal or non-seasonal, can be included in the model by suitable differencing schemes (Box and Jenkins, 1970). Models incorporating the dependence between the level and variance of a stochastic process can be developed by using various transformations, notably logarithmic and power transformations. The empirical autoregressive and autoregressive-moving averages models can be exteded to groups of species or species and environmental variables through the use of the multivariate analogues of the univariate models (Whittle, 1963). Finally, and perhaps most importantly, stochastic difference equation models can be created with dynamic characteristics, relating the fluctuations in a population to its past abundances, to the past values of uncontrollabe environmental variables, and to past values of other enivironmental variables which can be manipulated to some extent (Box and Jenkins, 1971; Astrom, 1970). These dynamic models can be used to predict changes in the abundance of the population given past observed values of the population and environmental variables and for the development of control schemes to maximize the density of the population, to minimize the abundance, or to hold the population fluctuations to some minimal variance given either feedback or feedforward control programs.

The most serious difficulty in the application of these models is gathering data on the population over a sufficiently long enough period of time to observe the characteristics of the population fluctuations and to acquire the sample sizes necessary to efficiently estimate the parameters of the model. In this study 28 observations over a period of two years and four months were available. By ecology's standards this is a long term study but by the statistician's the length of the series

of observations is marginal for efficient parameter estimation. However, studying a population long enough to observe the salient features of the fluctuations of the population is not a problem unique to stochastic difference equations or even to mathematical models in general. Population fluctuation is a long-term process. Short term studies will never yield enough information to understand or predict why populations fluctuate the way they do.

The only estimation procedures presented in this paper are the methods used to identify the appropriate order of the model and to find initial estimates of the parameters. The efficient least squares methods are too complex to discuss in this paper. The Yule-Walker parameter estimates of the autoregressive models appear, with one exception, to be fairly good. In practice, however, the efficient least-squares estimates should always be calculated after the preliminary identification and estimation of the model is completed. The computations are considerably more complex, of course, but if a computer program is available, the increased complexity is trivial. The reader is referred to the excellent discussion of the efficient estimation of the parameters of stochastic difference equations in Box and Jenkins (1970).

Models are valid for only those sets of conditions and interactions used in producing them. In unforseen events make drastic changes in the system, the models may become quite useless. The introduction or extinction of an important species of the community, drastic long term changes in the environment, the decision to build a dam or housing development, or a significant shift in the genetic makeup of the population may completely invalidate a model. Empirical models are useful tools, but like any other tool may become obsolete as conditions change.

In using these empirical models of population fluctuations, it is easy to lose track of why we have stipulated that the model should be stochastic and not deterministic. The rational behind stochastic models bares repeating. A deterministic model postulates some sort of functional relationship between a group of

variables. For example a variable $\underline{Y}$ is postulated to be a function of a set of $\underline{X}_i$'s. Given the $\underline{X}_i$'s, the function, and the parameters, the value of $\underline{Y}$ is exactly specified, usually by a single number. However, in the real world chance events and the influence of the almost infinite myriad of interacting environmental variables make the exact specification of the abundance of the population untenable philosophically and logically. By a purely subjective judgment the exactly specified prediction of the deterministic model may be considered to be some "best" or average prediction. This subjective leap implies a probabilistic interpretation of the prediction although this subjective definition of probabilities is impossible to completely interpret. We must accept the fact that it is impossible to exactly specify the future abundance of the population because of the stochastic nature of ecological processes such as population change. If an ecologist wishes to predict changes in population abundance, thes best he can do is to specify the probability distribution of the possible abundances of the population and choose some abundance as an "average","expected", or "most likely" outcome, where these terms have probabilistic interpretations. In other words the ecologist must "bet the odds" by choosing some group of possible outcomes and assigning a probability statement that the group will contain the true population abundance. This assignment of probabilities is exactly what a stochastic model attempts to do. The predicted abundance in forecasting should not be considered to be more than an average outcome. This average outcome is certainly convenient but perhaps less important than the probability statement assigned to a group of possible outocomes in the guise of a confidence interval. In fact we can specify the entire probability distribution of possible outcomes by selecting or assuming the distribution of the error terms of the model.

The emphasis on stochastic models is not meant to imply that stochastic models are real and deterministic models are not. Any model, stochastic or deterministic, is only a mathematical abstraction of something else. The difference between the models lies in the philosophy of prediction. Do we state with certainty what will

happen or do we talk about the chances of possible outcomes? The predictions of a stochastic model, although more humble than those of a deterministic model, at least recognize the importance of the imponderable conditions affecting mother nature.

Acknowledgments

I wish to thank Dr. Beverly J. Rathcke and Dr. Peter Feinsinger of Cornell University for several helpful comments. The research was partially supported by Public Health Training Grant 5-T01-GM-0392 of the National Institutes of Health to the Biometrics Unit at Cornell University.

Literature Cited

Anderson, T.W. 1971. The Statistical Analysis of Time Series. John Wiley and Sons. New York. 704pp.

Astrom, K.J. 1970. Introduction to Stochastic Control Theory. Academic Press. New York. 299pp.

Box, G.E.P., and G.M. Jenkins. 1970. Time Series Analysis: Forecasting and Control. Holden-Day, Inc. San Fransisco. 553pp.

Durbin, J. 1960. Estimation of parameters in time-series regression models. J. Royal Statist. Soc. Series B. 22:139-153.

Hannan, E.J. 1960. Time Series Analysis. Methuen and Co., Ltd. London. 152pp.

Hunter, A.S. 1966. High-altitude Drosophila of Colombia (Diptera: Drosophilidae). Annals Entomo. Soc. America 59:413-423.

Poole, R.W. 1972. An autoregressive model of population change in an experimental population of Daphnia magna. Oecologia 10:205-221.

Whittle, P. 1963. On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix. Biometrika 50:129-134.

Table 1. The autocorrelations and autocovariances of the three species of *Drosophila* for lags of zero to ten. The estimates are based on 28 observations.

	viracochi		mesophragmatica		pseudoobscura	
s	C_s	r_s	C_s	r_s	C_s	r_s
0	.1362	1.0000	.1719	1.0000	.1721	1.0000
1	.0539	.4282	.1225	.6996	.1035	.6526
2	.0391	.3268	.0678	.3737	.0911	.6417
3	.0694	.5590	.0315	.1673	.0412	.3101
4	.0120	.1123	-.0344	-.1911	.0701	.5415
5	.0070	.0666	-.0773	-.4280	.0678	.5171
6	.0016	.0146	-.0970	-.5146	.0809	.7152
7	-.0349	-.3086	-.1211	-.6153	.0512	.4606
8	-.0160	-.1389	-.0978	-.4874	.0292	.2950
9	-.0340	-.2832	-.0651	-.3160	.0158	.1532
10	-.0566	-.4631	-.0424	-.2071	.0342	.3311

Table 2. The estimated partial correlation functions of the three species of Drosophila.

	viracochi		mesophragmatica		pseudoobscura	
lag	ϕ_{pp}	$N^{\frac{1}{2}}\phi_{pp}$	ϕ_{pp}	$N^{\frac{1}{2}}\phi_{pp}$	ϕ_{pp}	$N^{\frac{1}{2}}\phi_{pp}$
1	.4282	2.2250	.6996	3.6352	.6526	3.3910
2	.1756	.8954	-.2267	-1.1559	.3759	1.9167
3	.4655	2.3275	.0107	.0535	-.3988	-1.9940
4	-.3910	-1.9155	-.5253	-2.5734	.6838	3.3500
5	-.0352	-.1688	-.0320	-.1055	.3131	1.5016
6	-.3781	-1.7734	-.2487	-1.1665	-.2295	-1.0765
7	-.1686	-.7726	-.2168	-.9935	-.2375	-1.0884
8	.1484	.6637	.1209	.5407	-.5923	-2.6488
9	-.1547	-.6743	-.3017	-1.3151	.8183	3.5669
10	.0002	.0008	-.0164	-.0696	2.3145	9.8196

Table 3. The variance-covariance matrices of the parameter estimates of the fourth order models applied to each of the three species of Drosophila. The standard errors of the estimates are the square roots of the diagonal elements of each matrix.

viracochi

.0303	-.0082	-.0003	-.0141
-.0082	.0259	-.0083	-.0003
-.0003	-.0083	.0259	-.0082
-.0141	-.0003	-.0082	.0303

mesophragmatica

.0259	-.0222	.0061	-.0003
-.0222	.0450	-.0274	.0061
.0061	-.0274	.0450	-.0222
-.0003	.0061	-.0222	.0259

pesudoobscura

.0190	-.0106	-.0102	.0076
-.0106	.0219	-.0008	-.0102
-.0102	-.0008	.0219	-.0106
.0076	-.0102	-.0106	.0190

Fig. 1. The fluctuations in log abundance of the three Drosophila species over 28 months at the Pine Woods site.

Fig. 2. Simulations of the fluctuations of mesophragmatica and pseudoobscura generated by the fourth order models applied to each species.

Fig. 3. The spectral densities of the three Drosophila species.

Fig. 4. The predicted log abundance of mesophragmatica for one to ten time intervals from time t. The open circles are the observed abundances and the bars 50% confidence intervals about the predicted abundance. The first four abundances are observed data and were used to generate the predictions.

Fig 1

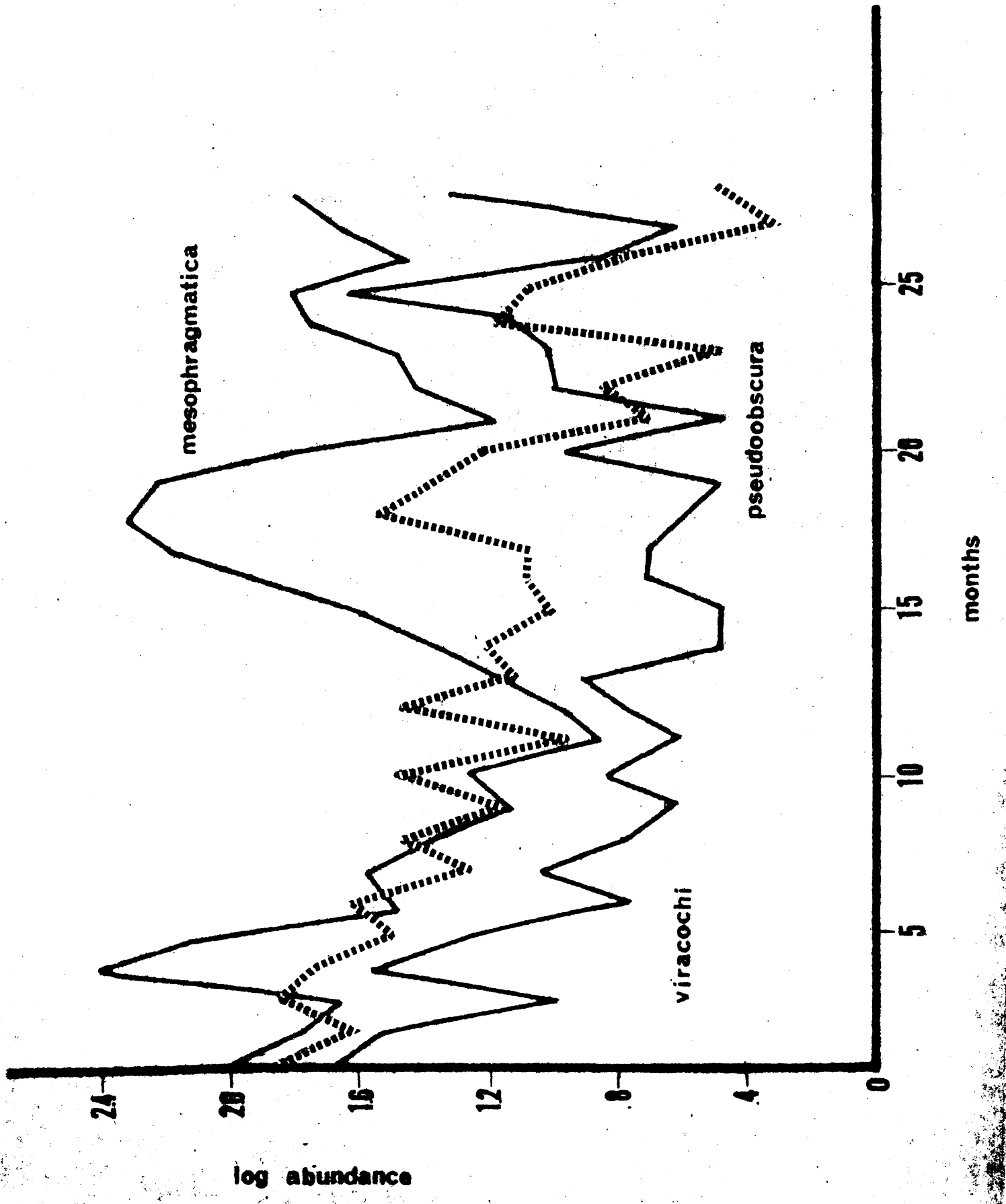
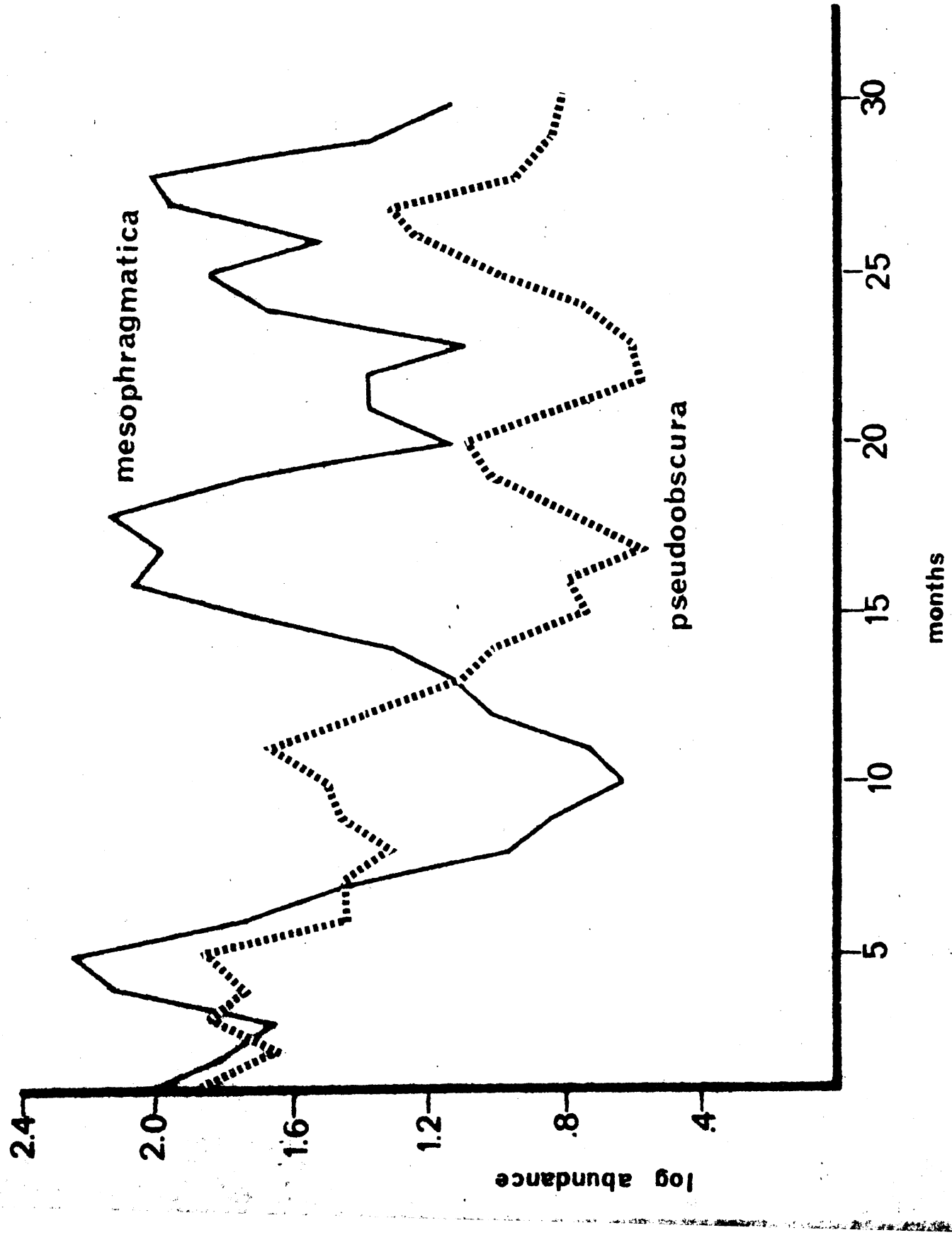

mesophragmatica
pseudoobscura
viracochi
log abundance
24
20
16
12
8
4
0
5
10
15
20
25
months

Fig 2


mesophragmatica
pseudoobscura
log abundance
2.4
2.0
1.6
1.2
.8
.4
5
10
15
20
25
30
months

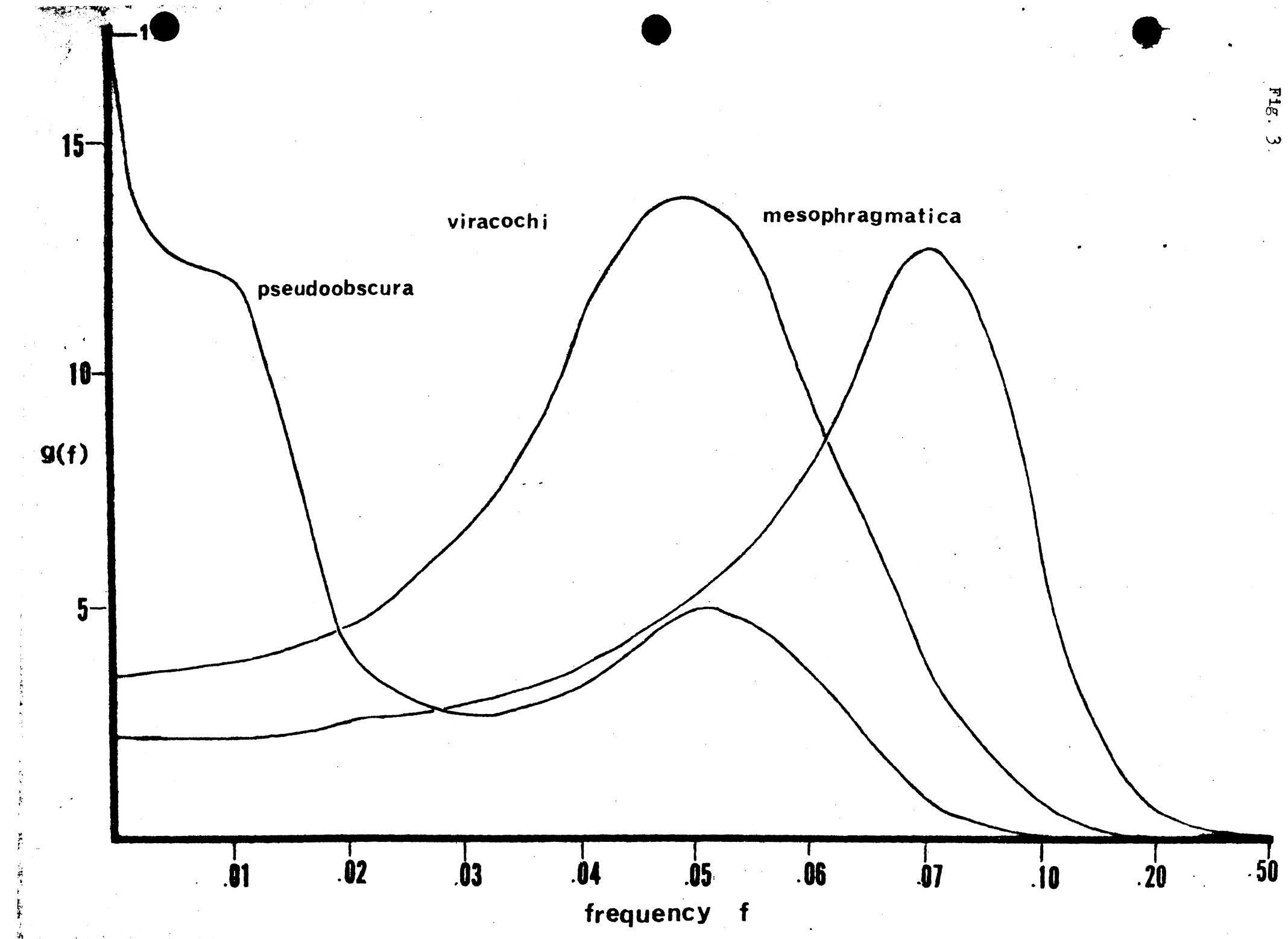

15
10
5
g(f)
pseudoobscura
viracochi
mesophragmatica
.01
.02
.03
.04
.05
.06
.07
.10
.20
.50
frequency f

Fig. 3

Fig. 4.